

Moving Large Data to Galaxy

Tom Doak

Le-Shin Wu

Carrie Ganote

National Center for Genome Analysis Support

July 16, 2014

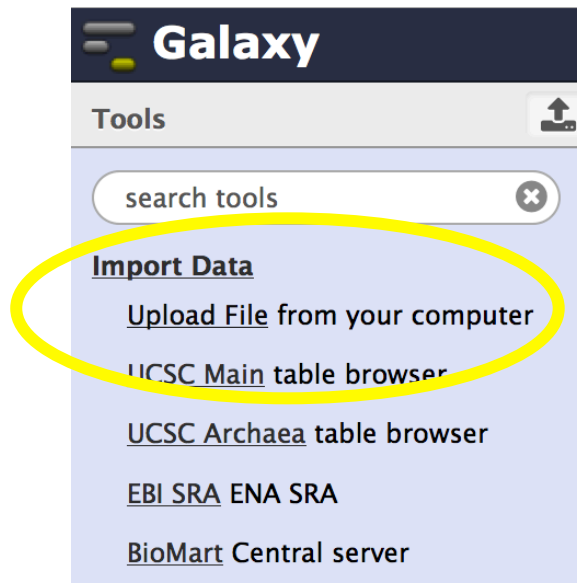


INDIANA UNIVERSITY



Let's get some sequence data

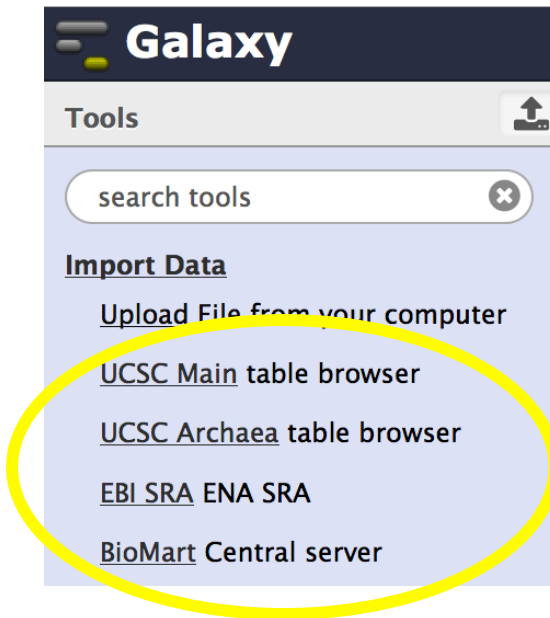
Moving large data sets onto Galaxy can be done in a number of ways:



For small files, use the Import Data -> Upload File from your computer. The page states a limit of 2GB, but in practice I wouldn't upload a file larger than a few MB this way - It's slow and tends to silently crash.



Let's get some sequence data



Data sources from UCSC and other browsers allow you to pull specific data from public resources



INDIANA UNIVERSITY

Let's get some sequence data

The best solution for large files is to move files directly to the server

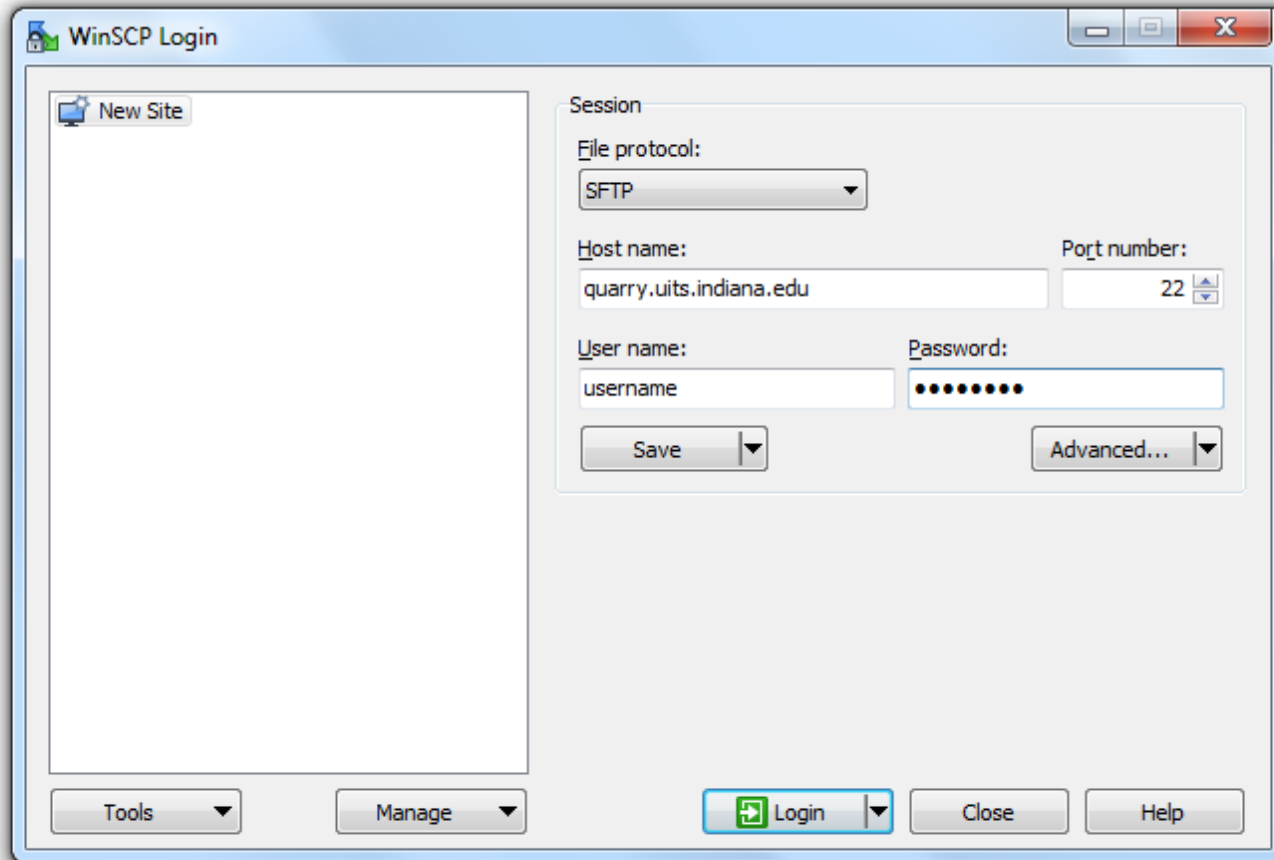
Any SFTP, SCP, or Globus client can be used to move files – we recommend WinSCP and Cyberduck

You will need an account on Mason, Quarry, or Big Red2. You can request accounts at itaccounts.iu.edu.



INDIANA UNIVERSITY

Let's get some sequence data



WinSCP

Login to host:

quarry.uits.indiana.edu

mason.indiana.edu

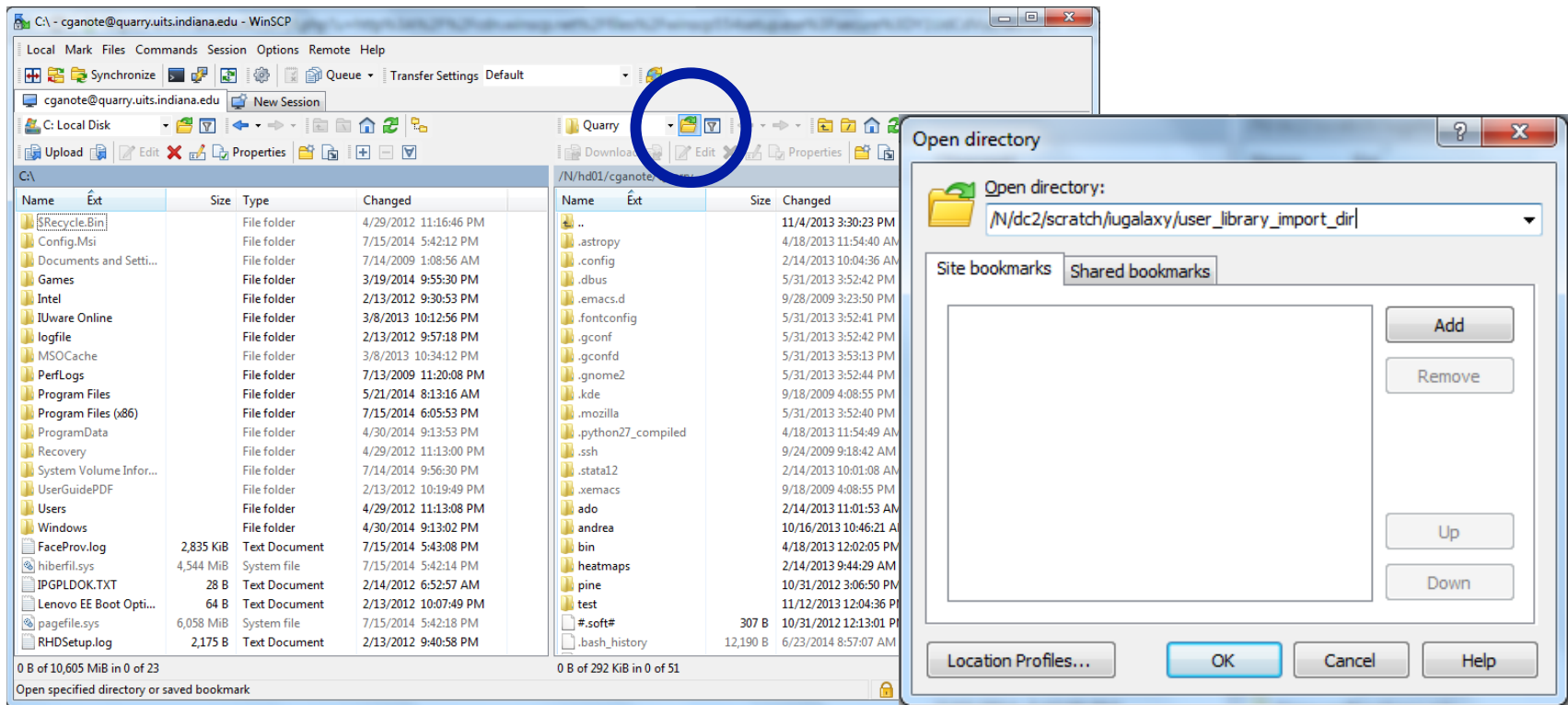
bigred2.iu.edu



INDIANA UNIVERSITY

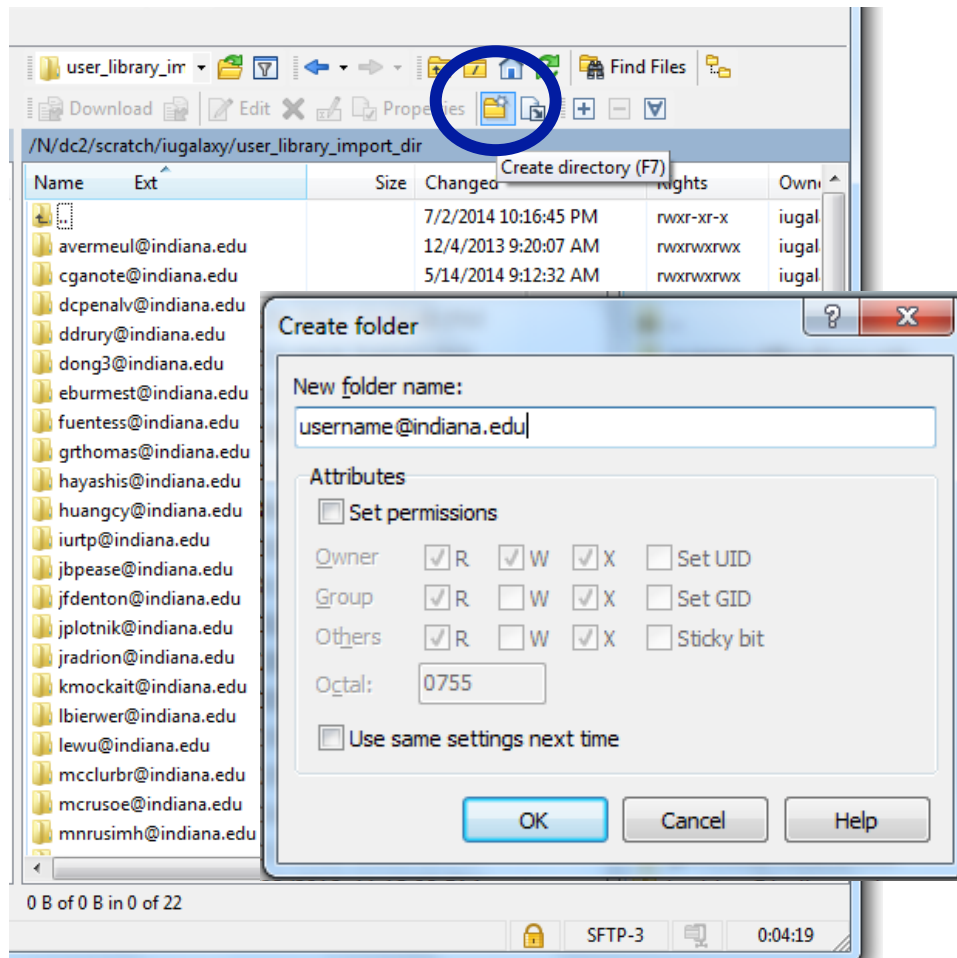
Let's get some sequence data

Open directory: /N/dc2/scratch/iugalaxy/user_library_import_dir



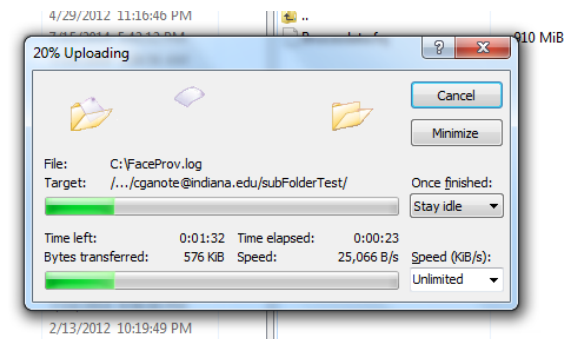


Let's get some sequence data



Create a folder if you don't have one: your username@indiana.edu

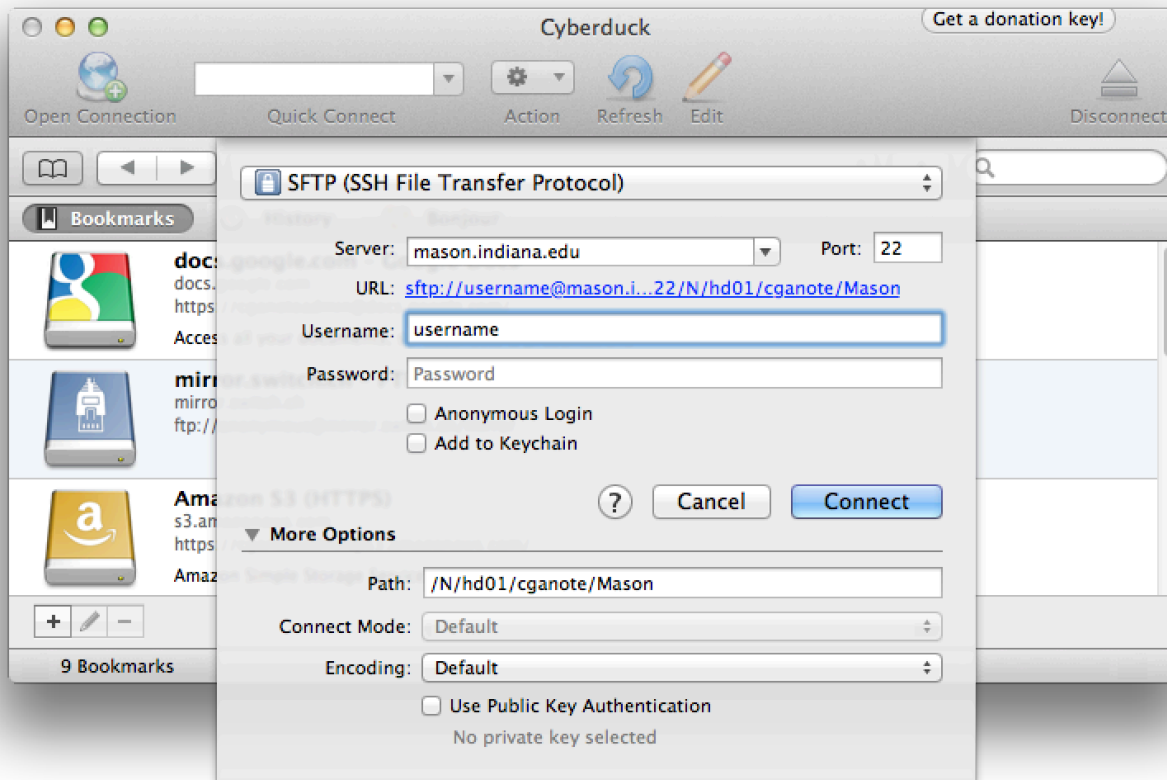
Then it is a matter of dragging and dropping files!





INDIANA UNIVERSITY

Let's get some sequence data



Cyberduck

Login to host:

quarry.uits.indiana.edu

mason.indiana.edu

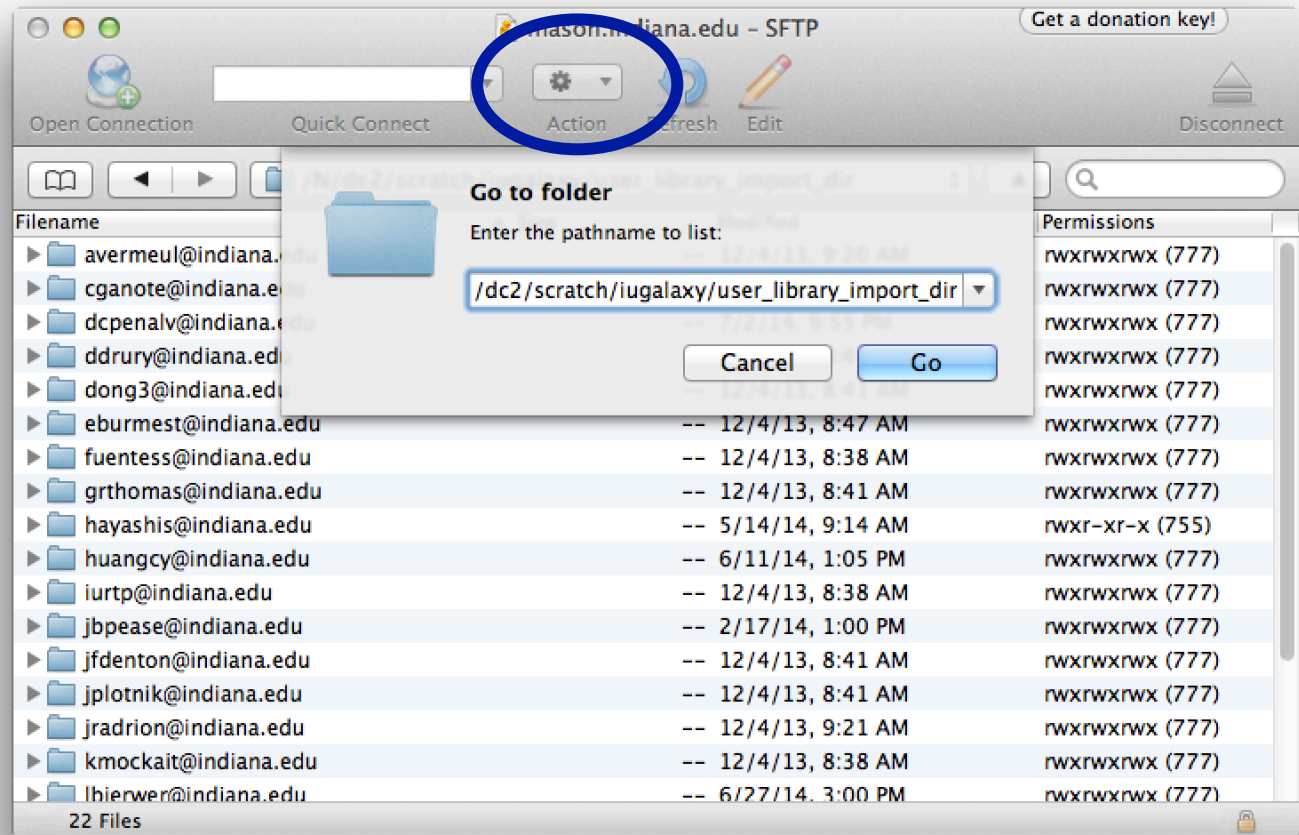
bigred2.iu.edu



INDIANA UNIVERSITY

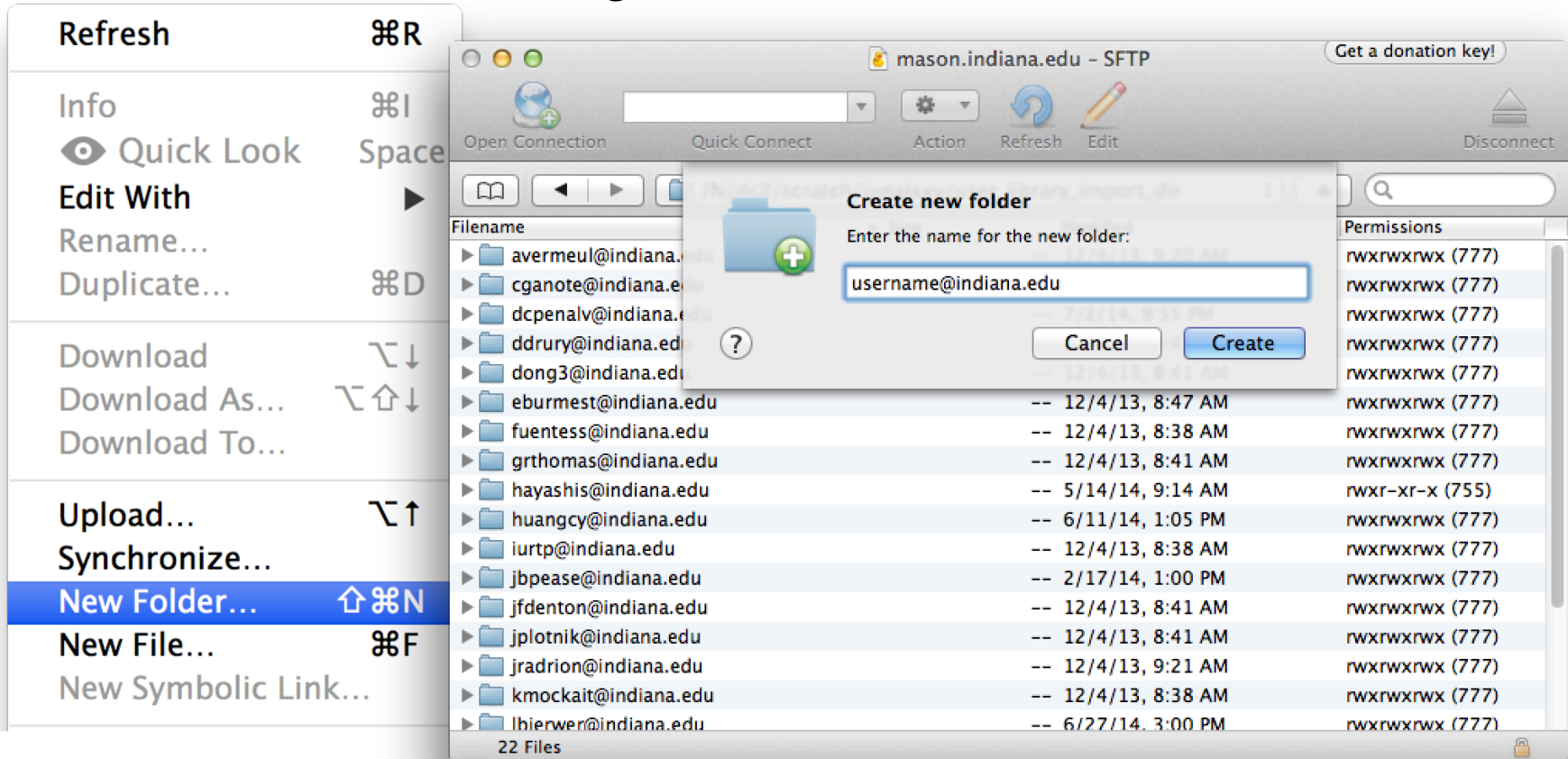
Let's get some sequence data

Actions:
Go to folder:
Same as before





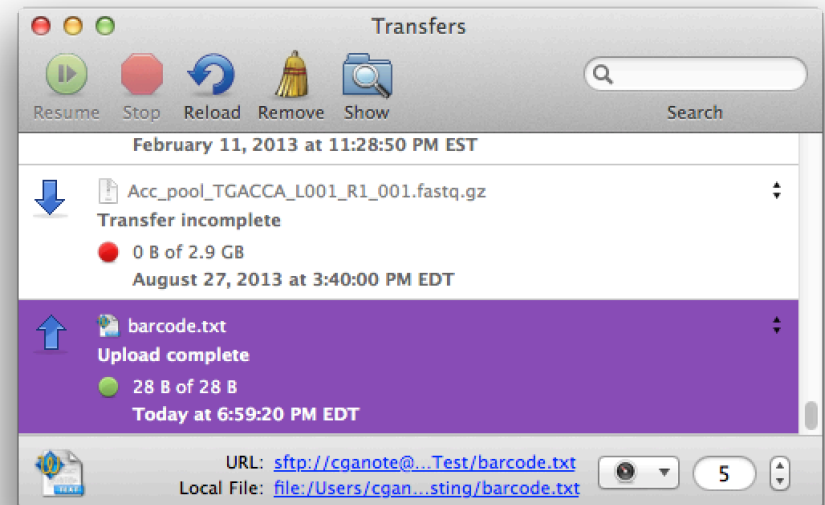
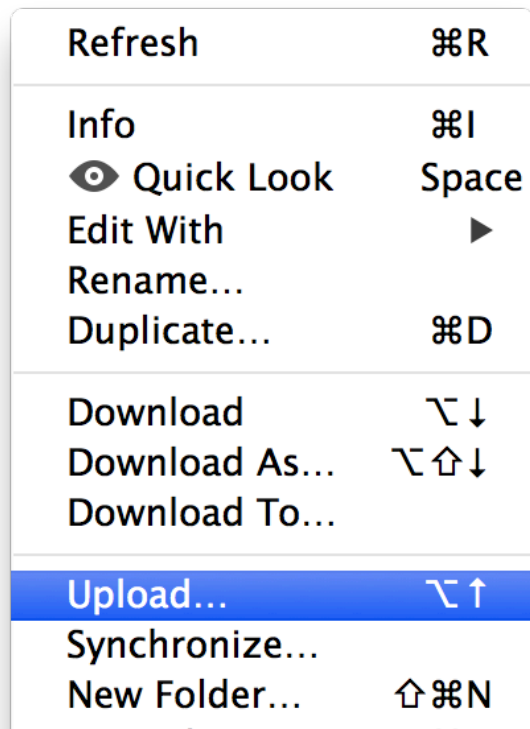
Right click to create a new folder





Let's get some sequence data

Cyberduck only shows the remote view, so when you are in the folder you want, right click and choose Upload to browse local files





Let's get some sequence data

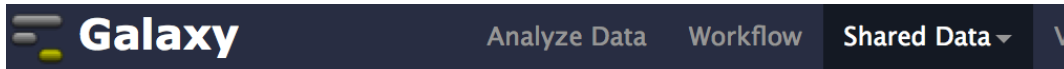
The screenshot shows the Galaxy web interface. At the top is a navigation bar with the 'Galaxy' logo and links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Admin', and 'Help'. Below this is the 'Data Libraries' section, which includes a search bar with the placeholder text 'search dataset name, info, message, dbkey' and a magnifying glass icon. Below the search bar is a link for 'Advanced Search'. A table of data libraries is displayed below, with two columns: 'Data library name ↓' and 'Data library description'. The first row in the table is 'User Import Library', which is circled in blue, with the description 'For moving large datasets into Galaxy'. The second row is 'Workshop Data' with the description 'Learning sets of RNA-Seq data'.

Data library name ↓	Data library description
User Import Library	For moving large datasets into Galaxy
Workshop Data	Learning sets of RNA-Seq data






Go to Shared Data -> Data Libraries -> User Import Library



Let's get some sequence data



Data Library "User Import Library"

<input type="checkbox"/> Name	Message	Data type	Date
<input type="checkbox"/>  avermeul@indiana.edu	Arjan's Files		
<input type="checkbox"/>  cganote@indiana.edu			
<input type="checkbox"/>  dcpenalv@indiana.edu			
<input type="checkbox"/>  eburmest@indiana.edu	Liz's Files		
<input type="checkbox"/>  hayashis@indiana.edu	Soichi's Files		

Find your username – click on the black arrow to right of it

If you don't see your name, contact us at help@ncgas.org and we will add you ASAP!



Let's get some sequence data

Galaxy Analyze Data Workflow Shared Data ▼

Data Library "User Import Library"

<input type="checkbox"/>	Name	Message	Data type	Date uploaded
<input type="checkbox"/>	avermeul@indiana.edu ▼	Arjan's Files		
<input type="checkbox"/>	cganote@indiana.edu ▼			
<input type="checkbox"/>	[dropdown]			
<input type="checkbox"/>	[dropdown]			
<input type="checkbox"/>	[dropdown]			
<input type="checkbox"/>	[dropdown]			
<input type="checkbox"/>	[dropdown]			
<input type="checkbox"/>	[dropdown]			
<input type="checkbox"/>	[dropdown]			
<input type="checkbox"/>	[dropdown]			

- Add datasets
- Add sub-folder**
- Select datasets for import into selected histories
- Edit information
- Move this folder
- Use template
- Make public
- Edit permissions
- Delete this folder

Add a sub-folder
with a descriptive
name

Create a new folder

Name:

Description:



Let's get some sequence data

Galaxy Analyze Data Workflow Shared Data View

Data Library "User Import Library"

✓ The new folder named 'A descriptive name' has been added to the data library.

Name	Message
avermeul@indiana.edu	Arjan's Files
cganote@indiana.edu	
A descriptive name	very descriptive
Bruce's	
SRA	
Test2	Testing subfolder uploads
Test da	Carrie's test data

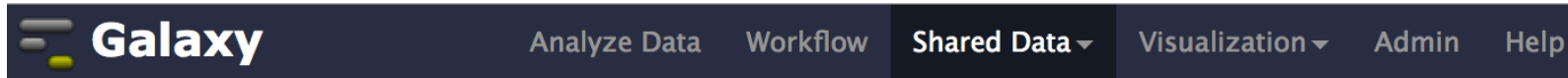
- Add datasets
- Add sub-folder
- Edit information
- Move this folder
- Use template
- Edit permissions
- Delete this folder

Right click on the black arrow to the right of the new folder.

Choose to Add datasets.



Let's get some sequence data



Upload files to a data library

Browse this data library

Upload a directory of files

Upload option:

Upload directory of files

Choose upload option (file, directory, filesystem paths, current history).

File Format:

Auto-detect

Server Directory

SRA

Upload all files in a sub-directory of `/N/dc2/scratch/iugalaxy/user_library_import_dir/cganote@indiana.edu` on the Galaxy server.

Copy data into Galaxy?

Link to files without copying into Galaxy

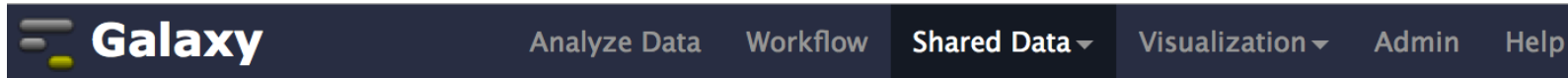
Normally data uploaded with this tool is copied into Galaxy's configured "file_path" location where Galaxy has a

Always upload a directory of files.

If you ever forget where to put your files, the directory is listed here:



Let's get some sequence data



Upload files to a data library

Browse this data library

Upload a directory of files

Upload option:

Upload directory of files

Choose upload option (file, directory, filesystem paths, current history).

File Format:

Auto-detect

Server Directory

SRA

Upload all files in a sub-directory of `/N/dc2/scratch/iugalaxy/user_library_import_dir/cganote@indiana.edu` on the Galaxy server.

Copy data into Galaxy?

Link to files without copying into Galaxy

Normally data uploaded with this tool is copied into Galaxy's configured "file_path" location where Galaxy has a

Up to one subdirectory on the server will be recognized

Link large files, but make sure they are backed up!



Let's get some sequence data

Galaxy

Analyze DataWorkflowShared DataVisualizationAdminHelpUser

Data Library "User Import Library"

✓ Added 3 datasets to the folder 'A descriptive name' (each is selected). Click the Go button at the bottom of this page to edit the permissions on these datasets if necessary.

<input type="checkbox"/>	Name	Message	Data type	Date uploaded	File size
<input type="checkbox"/>	avermeul@indiana.edu ▾	Arjan's Files			
<input type="checkbox"/>	cganote@indiana.edu ▾				
<input type="checkbox"/>	A descriptive name ▾	very descriptive			
<input checked="" type="checkbox"/>	barcode.txt ▾		auto	Wed Jul 16 02:44:12 2014 (UTC)	28 bytes
<input checked="" type="checkbox"/>	Brucesdata.fq ▾		auto	Wed Jul 16 02:44:13 2014 (UTC)	909.9 MB
<input checked="" type="checkbox"/>	FaceProv.log.filepart ▾		auto	Wed Jul 16 02:44:13 2014 (UTC)	1.4 MB

Import to current history
Import to histories
Edit permissions
Move
Delete
Download as a .tar.gz file
Download as a .tar.bz2 file
Download as a .zip file

For selected datasets:

Import to current history ▾

Go



INDIANA UNIVERSITY

Fin

Thanks for watching!
Questions and comments:
Email help@ncgas.org